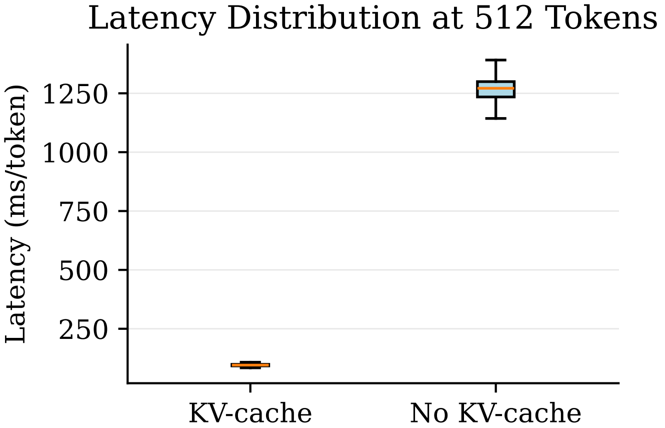
Latency Distribution at 512 Tokens
Latency (ms/token)
1250
1000
750
500
250
KV-cache
No KV-cache